RNA-Seq Demo on Galaxy

Tom Doak
Le-Shin Wu
Carrie Ganote
National Center for Genome Analysis Support

July 16, 2014

INDIANA UNIVERSITY

Our RNA-Seq Demo Data

Cristobal Rojas, La miseria (1886)

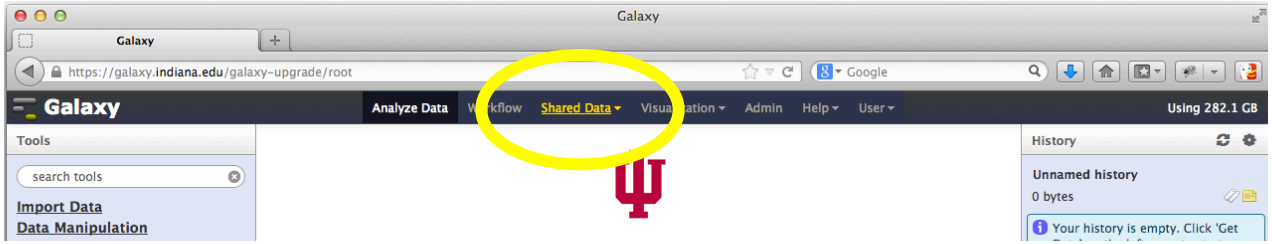
We will be assembling the DNA Polymerase protein units from the H37Rv strain of *Mycobacterium tuberculosis*, the causative agent of TB, also known as the consumption.

The raw reads originated from the Short Read Archive on NCBI. The accession number for the set is SRX212035.

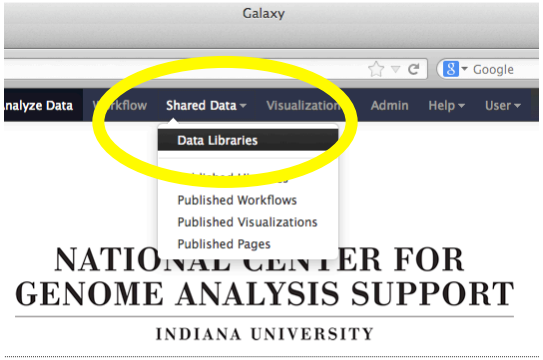
This dataset consists of paired-end, ~75bp RNA-Seq reads.

Let's get some sequence data

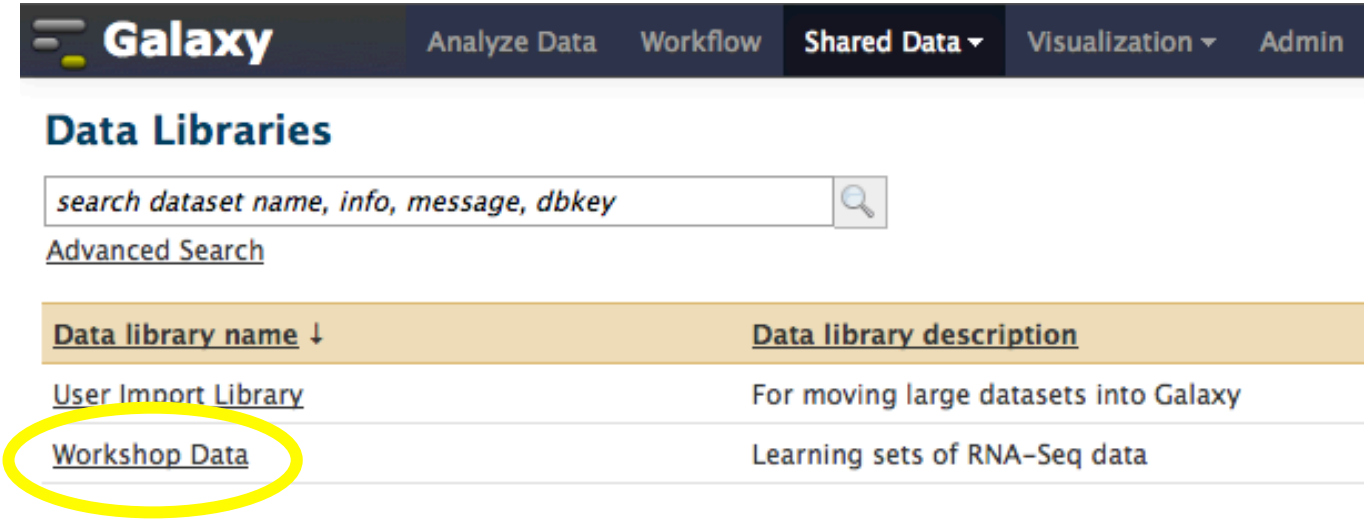
Galaxy allows users to publish their data to share with each other.



Let's start with "Shared Data" at the top.
Then select Data Libraries from the menu.



Let's get some sequence data

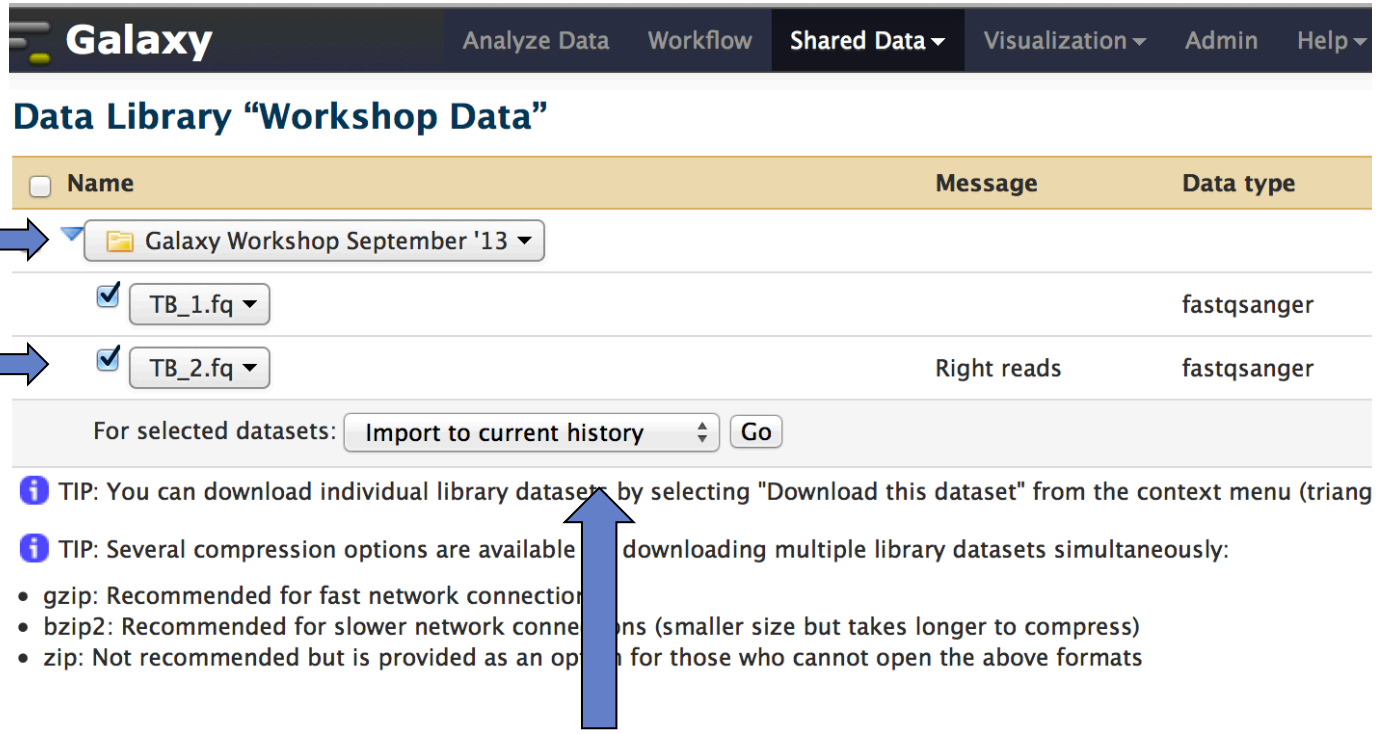


Choose Workshop Data.

Let's get some sequence data

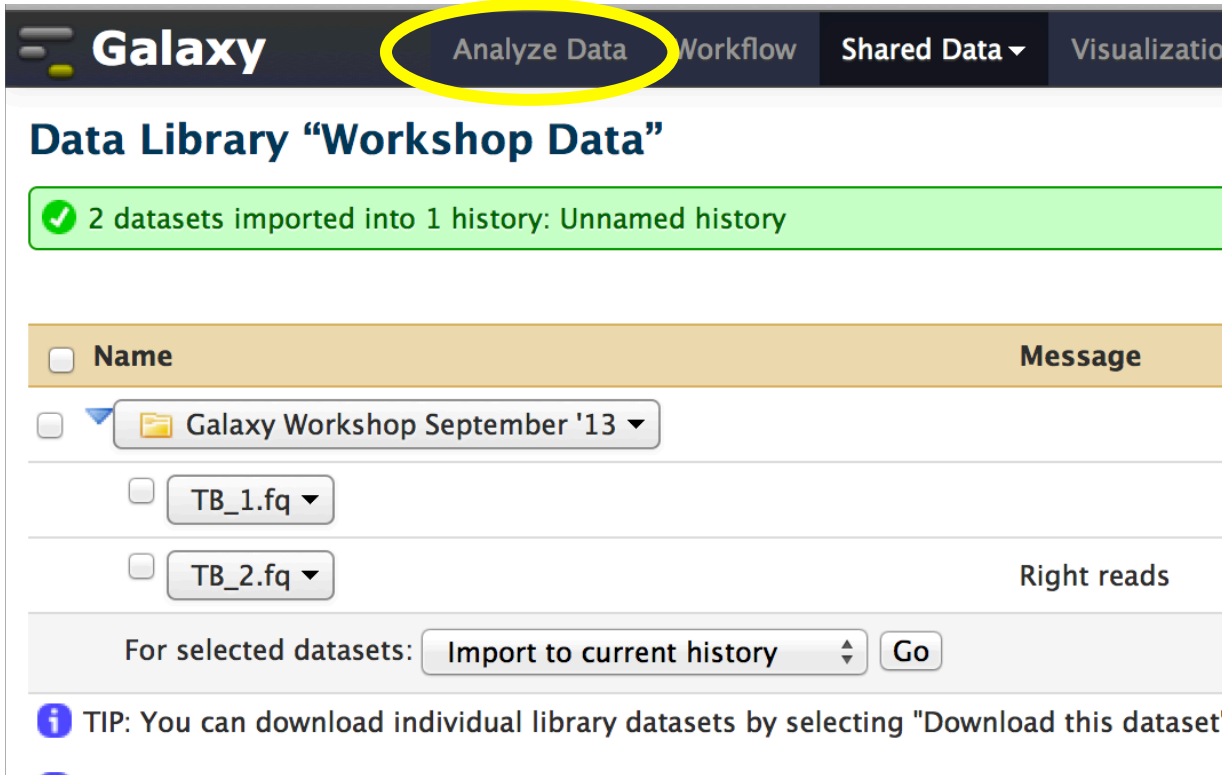
Expand folder

Check both boxes



Import the Data sets to current history.

Let's get some sequence data

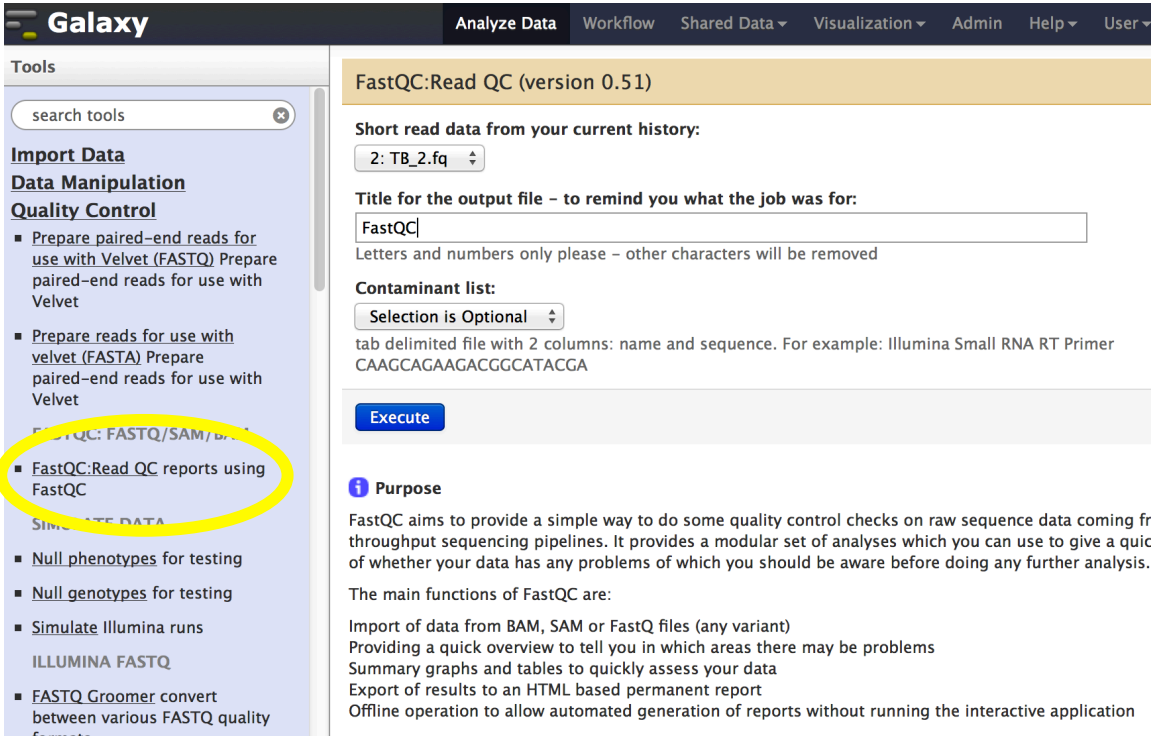


Data set is imported – Click on Analyze Data to return.

Step 1: Assess the Quality of Inputs

We will first get an idea of the quality of our input data sets.

The FastQC tool will produce graphical output that makes it easy to gauge the characteristics of the data – quality, patterns, biases, gc content etc.

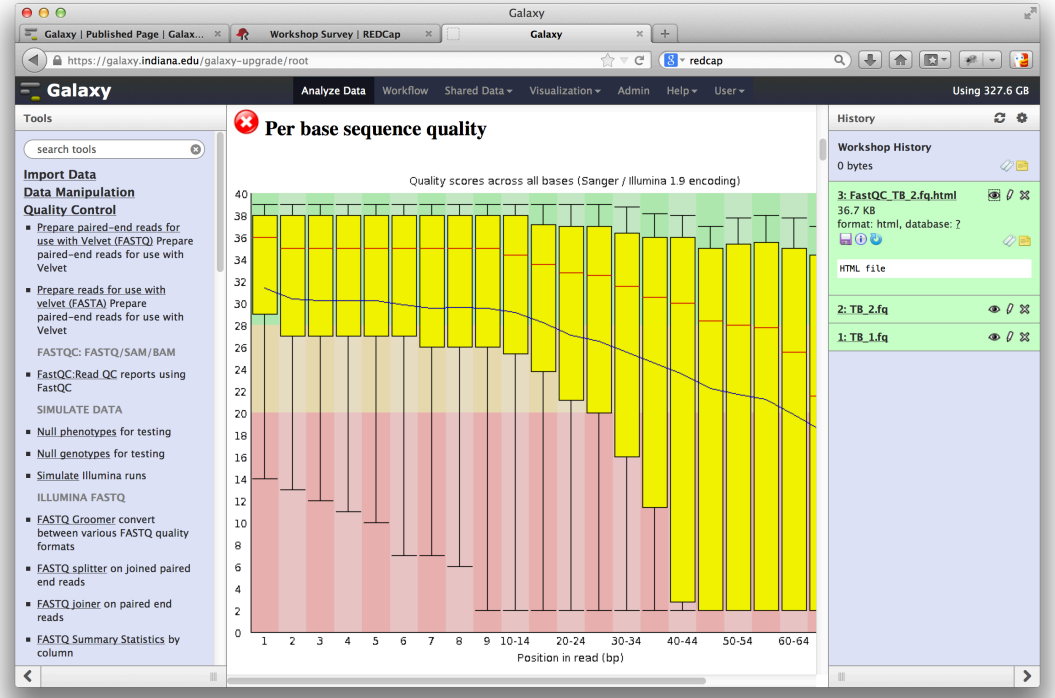


Choose either the left or right reads. Compare the results with your neighbor.

Step 1: Assess the Quality of Inputs

The input data usually declines in quality as the reads progress.

The quality score is assigned by the sequencing machine as it reads each base. It is a rough estimate of how ambiguous the signal is.



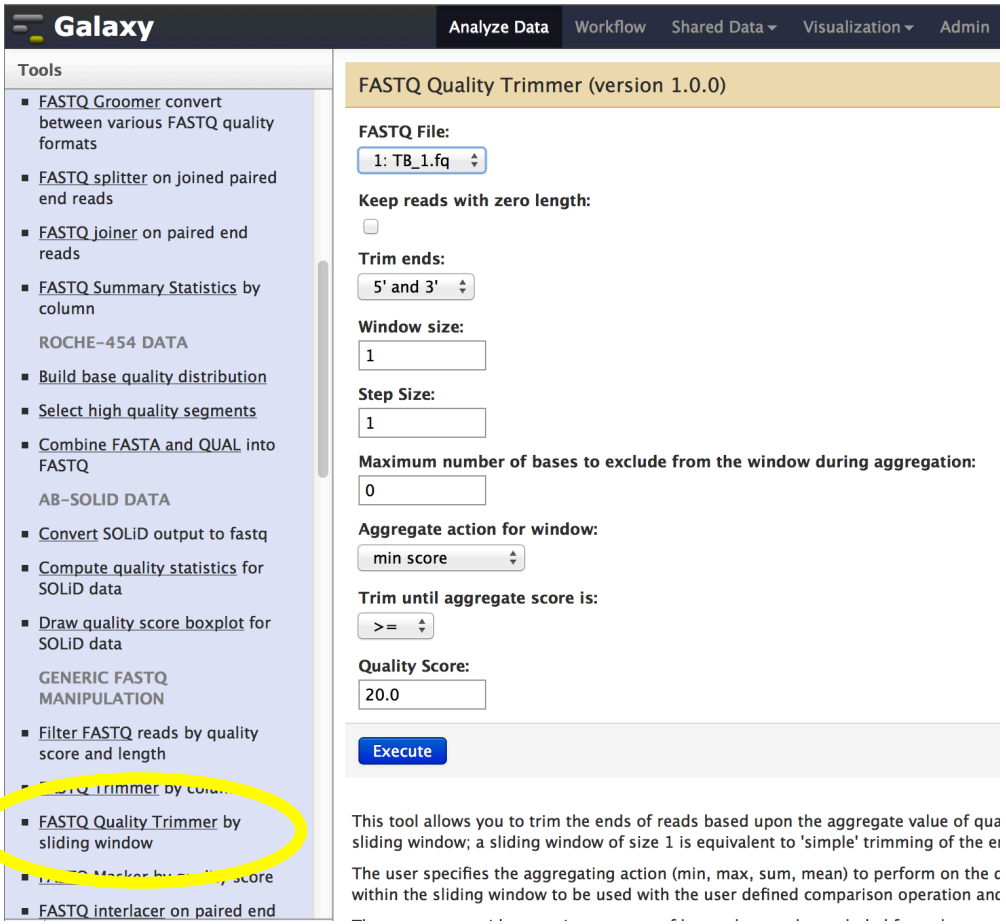
Sequence: ATGCATG

Quality Score: 39 38 23 19 3 3

Step 2: Trim Input Sequences

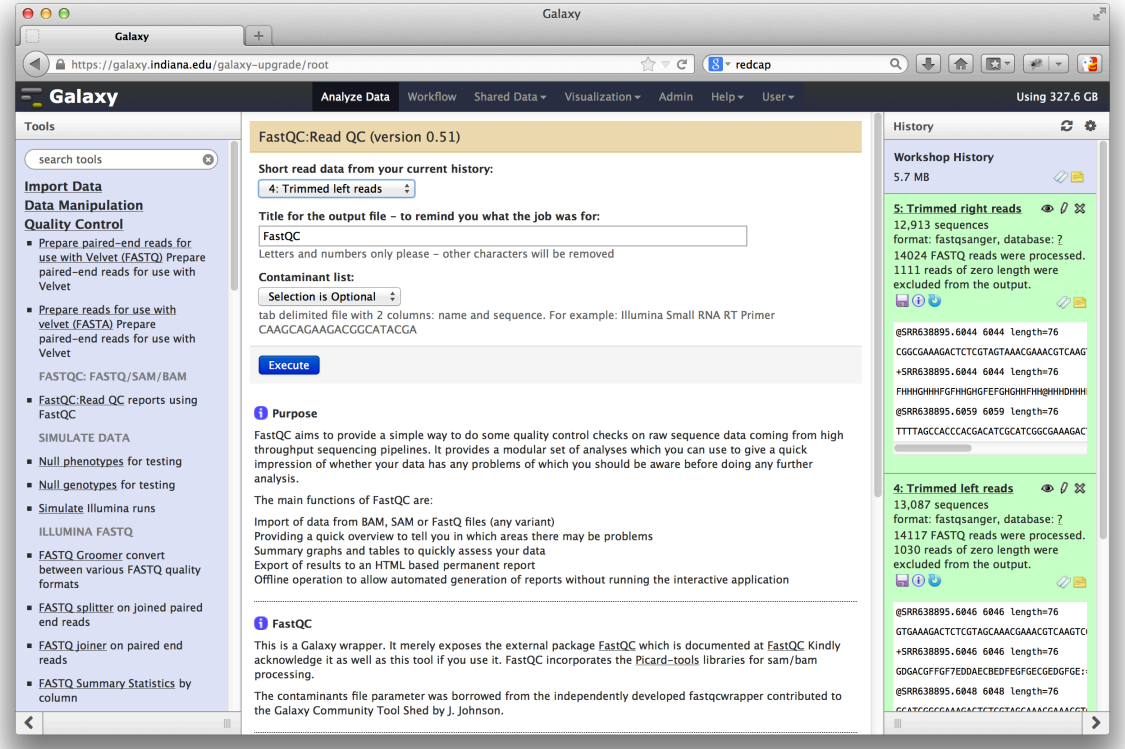
We've determined that the input data sets need some work before they are used in downstream processes. We'll use the FASTQ quality trimmer by sliding window to trim reads based on quality score.

Run this tool for both input data sets.



Step 3: Rinse, Repeat

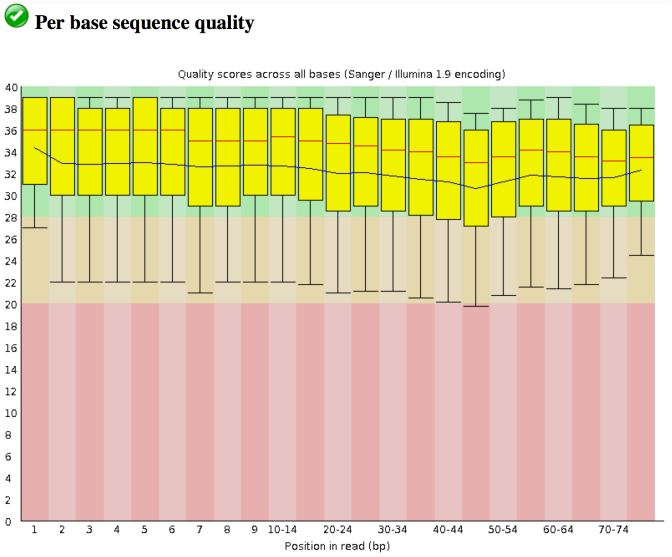
Now that the files are trimmed, we will re-assess their quality. If necessary, keep trimming away until you are satisfied with the input files.

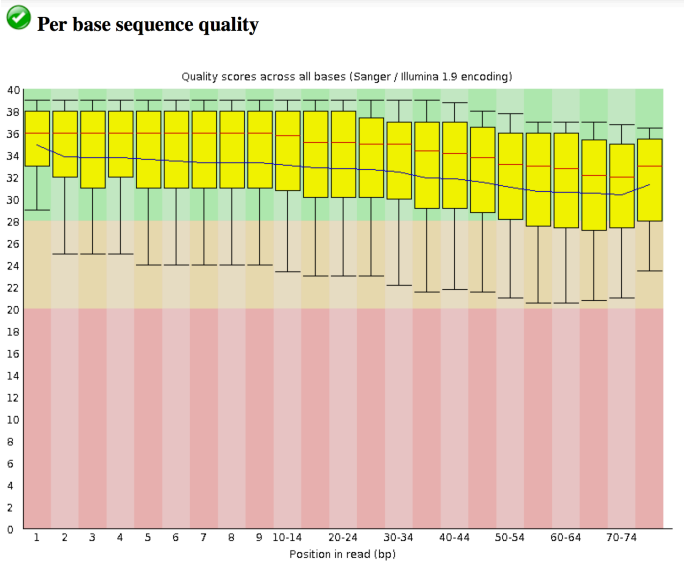


I renamed my trimmed files to help me keep them straight.

Step 3: Rinse, Repeat

Pictured are the left and right reads after trimming is complete.
These will do!

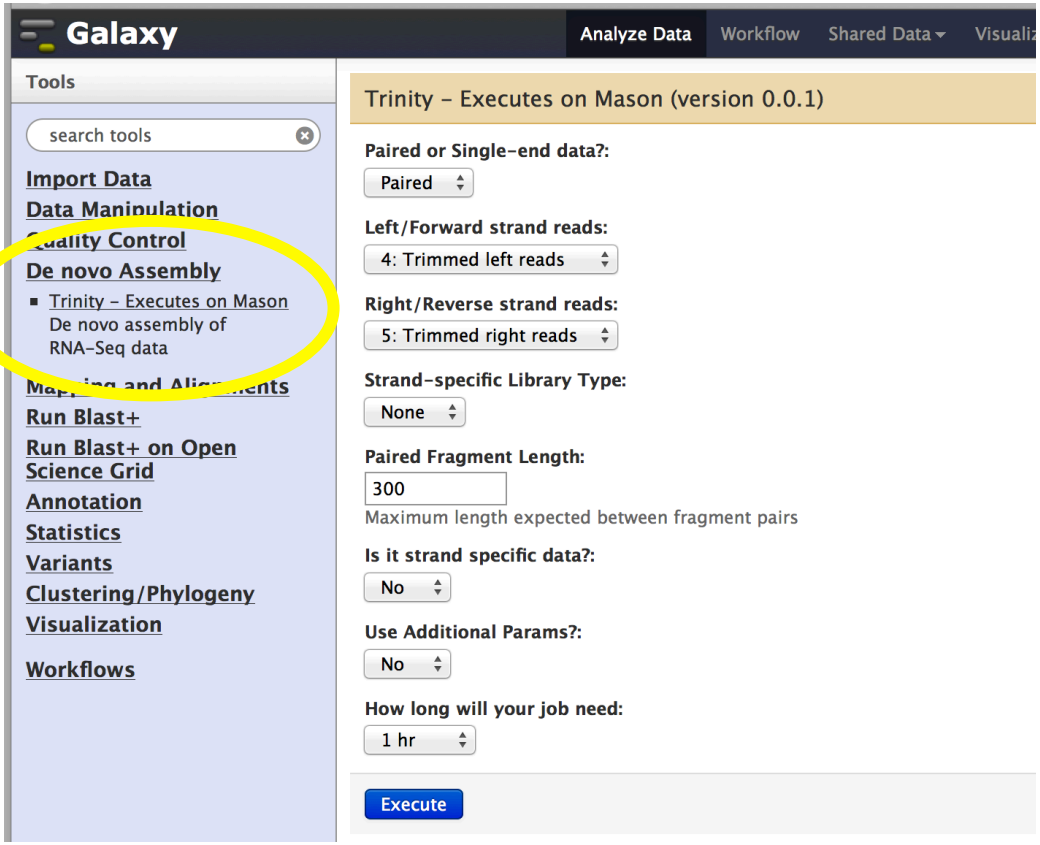




Step 4: Assembly

Next we will put the reads together to create a complete picture of the actively transcribed genes of the sample organism.

Trinity is a *de novo* assembler that has been optimized for use on Mason. We will use it to assemble our reads.



It finished! We're done, right?

An assembler solves a computer problem of putting together a puzzle from tiny pieces. The output of the assembler is a guess – but we don't know how accurate it is. We could look at:

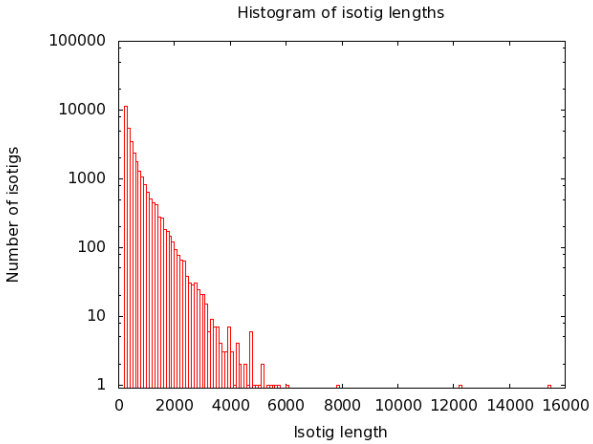
- Basic stats of the assembly – "Contigs"
  - Number of "Contigs" vs. Expected Number
  - N50 – a weighted average
  - Average Length
  - Max Length
- Check contigs against known genes with Blast

Step 5: Assessing Quality of Assembly

Important statistics for assembly quality:

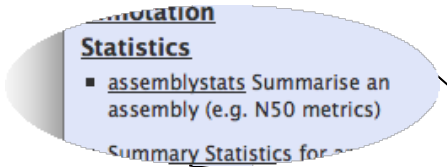
Contig Length Distribution

Assemblies will typically produce a number of complete contigs representing whole transcripts, and a large number of partial transcripts. This biases the average contig length toward the low end. The N50 is a measure weighted by total sequence length in the assembly.

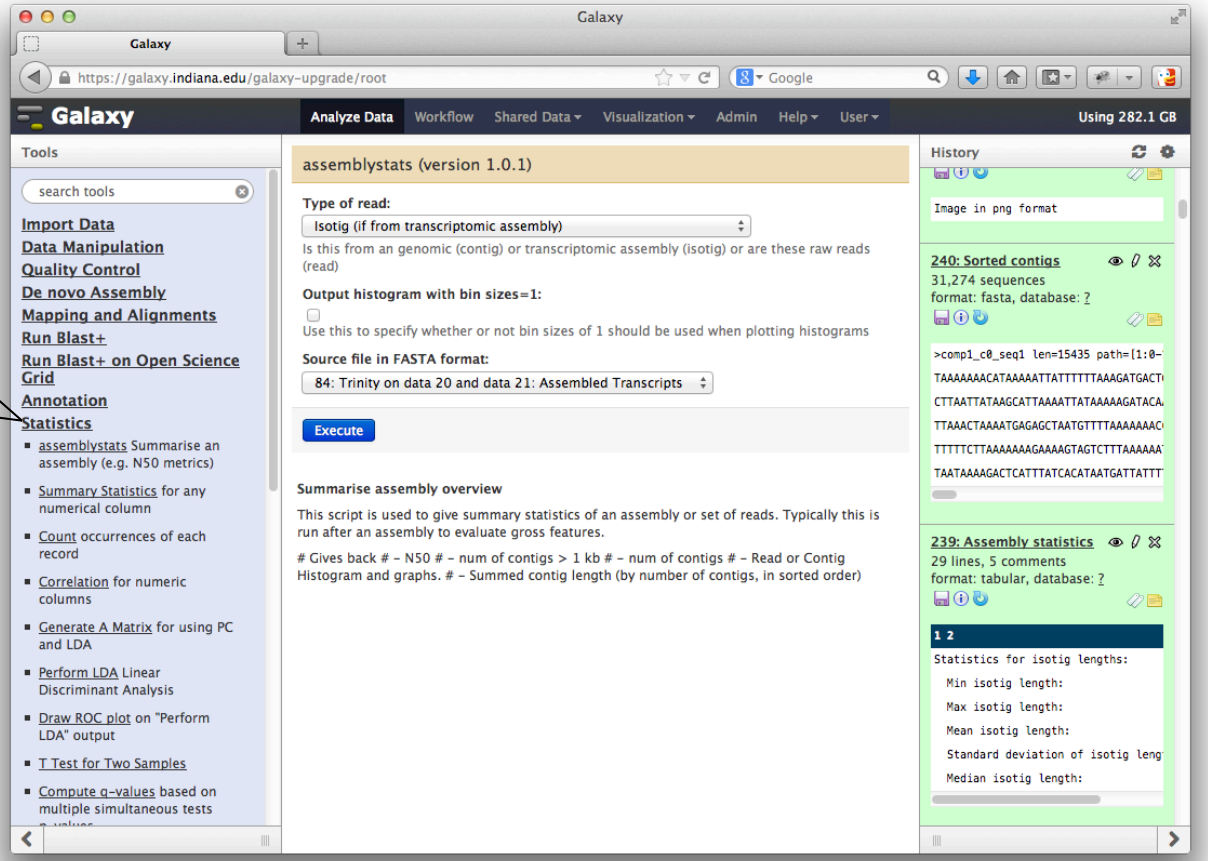


Step 5: Assessing Quality of Assembly

Getting these stats in Galaxy:



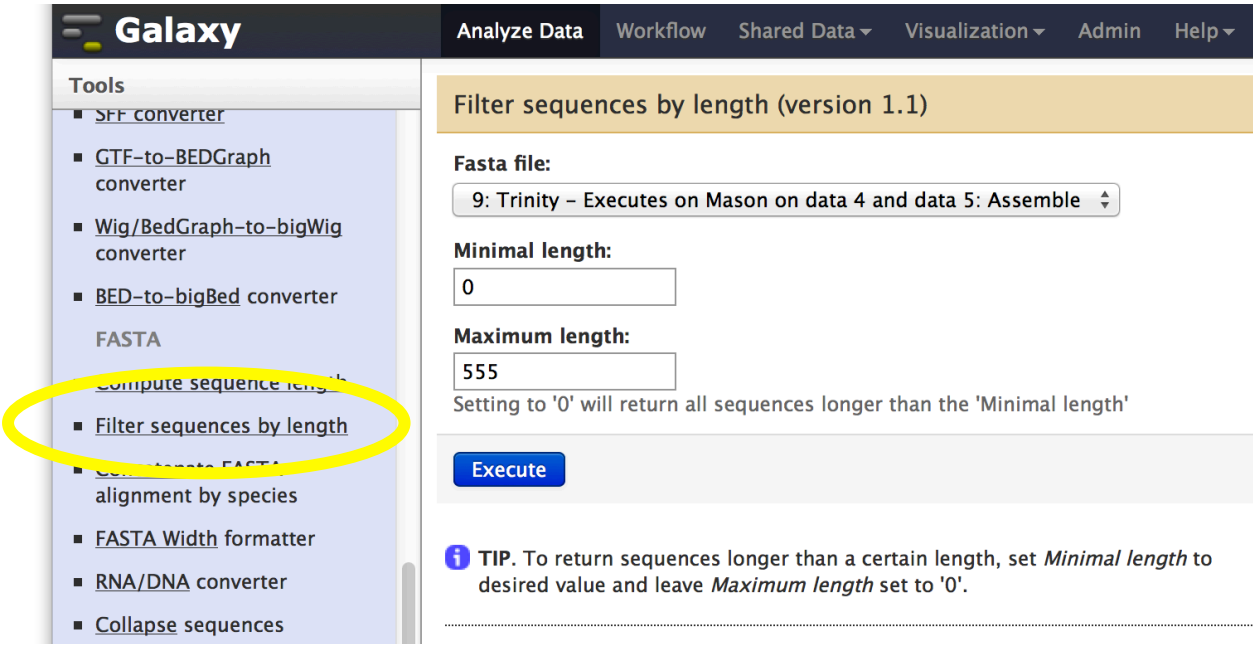
Run assemblystats to get a summary and histograms of your contig length distribution.



Step 6: Check Against Database

For this last step, we'll check to see how well our assembled transcripts compare to what we already know.

Use this step to give a rough annotation of genes, to make sure that your transcripts are from nuclear genes, or to gauge how complete your sequence is.

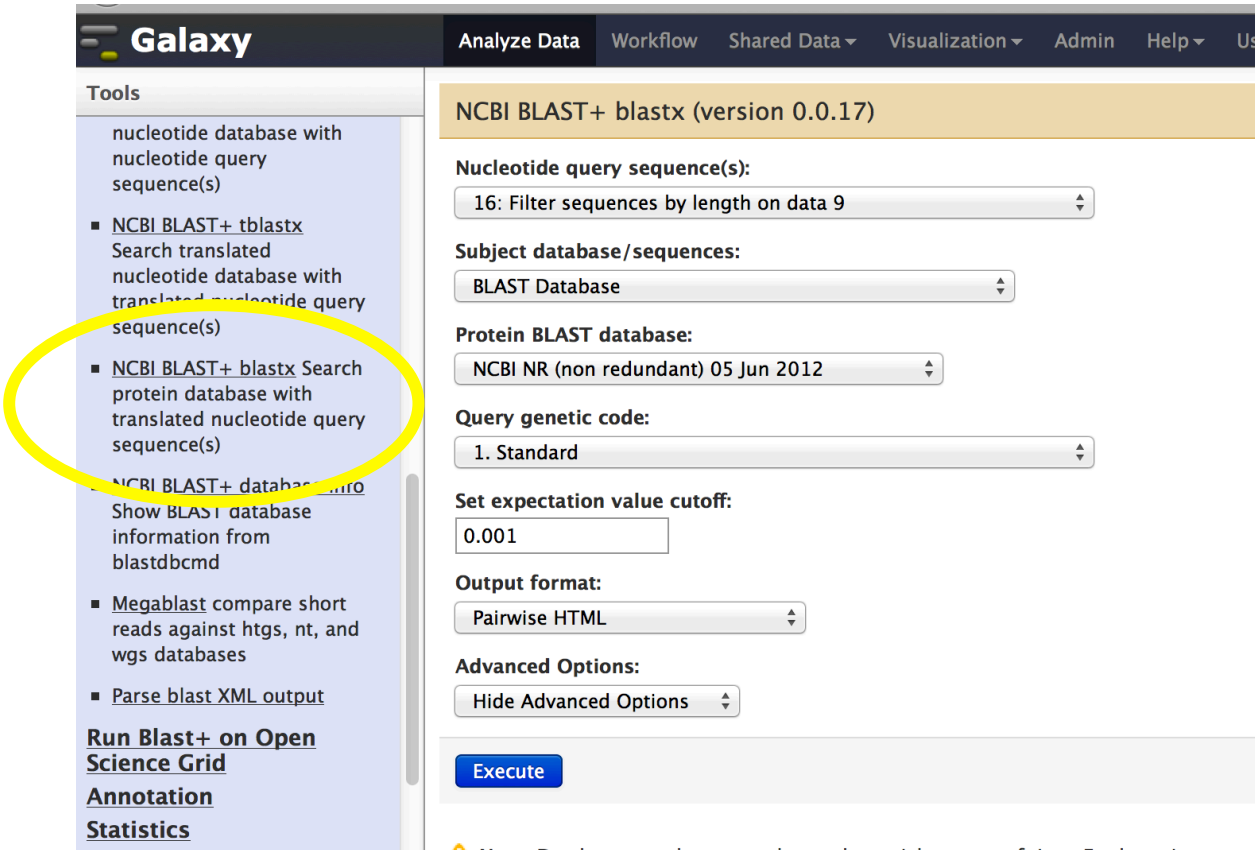


For sake of time, we'll just Blast one gene. Filter out to get the smallest.

Step 6: Check Against Database

We will use Blastx to search the NR database for our gene.

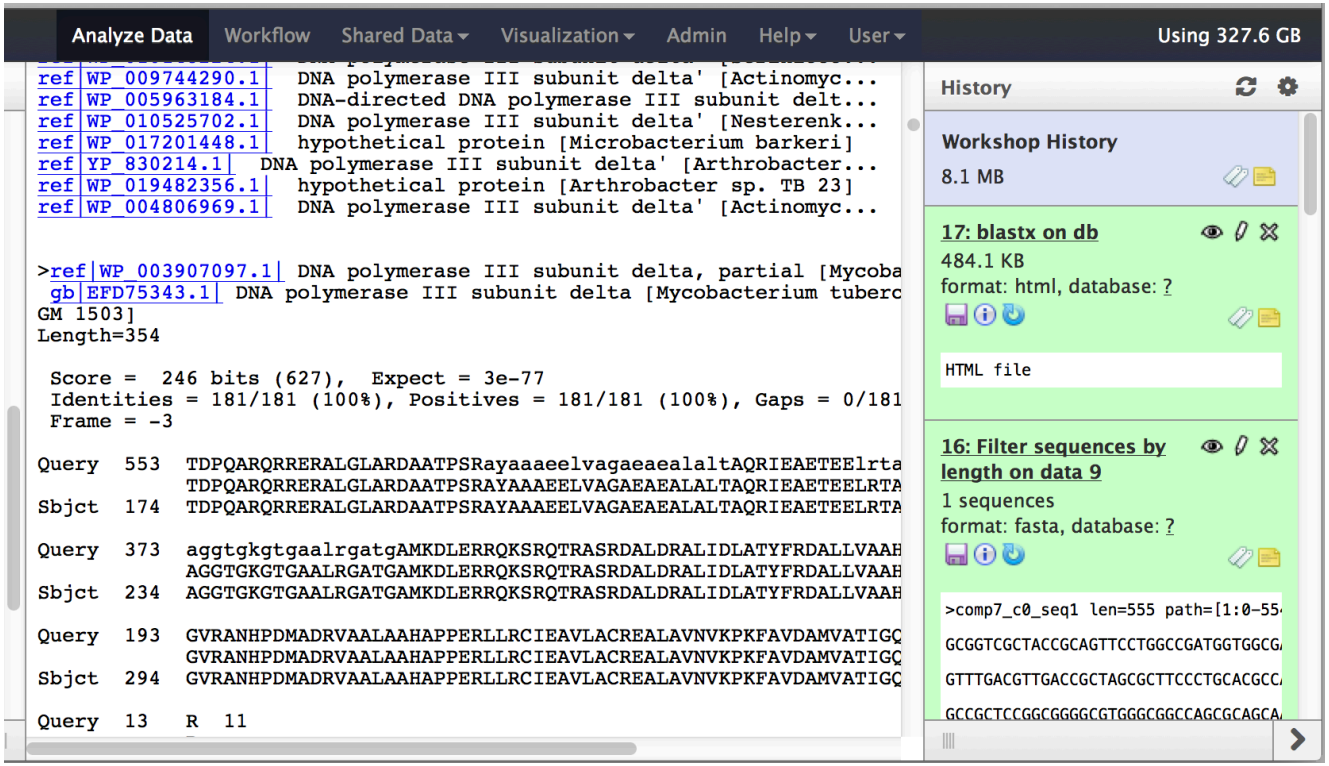
Use default search settings for this test set.



Make sure to choose Pairwise HTML output for readability.

Step 6: Check Against Database

We see the expected genes as the top hits!



We could limit the number of hits depending on output desired.

Step 7..?

RNA-Seq is a very versatile technology. You can use the data for:

- Gene discovery based on transcripts
- Genome evidence – introns, exons, junction
- Gene expression patterns
- SNP calling/other variants
- Protein divergence between samples

We have gotten to the assembly step, but there is a lot to learn about the data now that it is put together. A foundation in the use of Galaxy coupled with Indiana University resources will enable you to reach these goals.

Fín

Thanks for watching!

Questions and comments:

Email help@ncgas.org